CAICT
中国信通院

HUAWEI

数据基础设施白皮书

2019

目录

前言

人类社会几千年来经历了农业经济、工业经济，如今已经进入到数字经济时代。根据联合国《2019年数字经济报告》的统计，数字经济的规模估计占全球生产总值的4.5%至15.5%之间，其中中国和美国是引领世界数字经济发展的核心。《中国互联网发展报告2019》指出，2018年，中国数字经济规模达31.3万亿元，占GDP比重达34.8%，数字经济已成为中国经济增长的新引擎，正在深刻改变全社会的生产和生活方式。

虽然学界对数字经济的构成模式和理论体系还没有清晰的界定，但数据作为数字经济时代最有价值的生产资料已经是毋庸置疑的共识。云计算、大数据、物联网、移动互联网、人工智能等ICT新技术、新模式的发展和应用无一不是以海量数据为基础，又反过来带动了数据量的爆发式增长。

就像石油的“采-运-炼-储-用”是工业经济的核心命脉一样，面向海量数据的“采-存-算-管-用”是支撑数字经济运行的基础能力。海量数据蕴含巨大的价值，也带来了前所未有的挑战，数据“存不下、流不动、用不好”成为了各行业数据应用最普遍的难题，以“融合、协同、智能、安全、开放”为特征的新型数据基础设施可以帮助各行业实现数据存储智能化、管理简单化和价值最大化，是推动各行业拥抱数字经济浪潮的关键因素之一。

在此背景下，中国信息通信研究院和华为技术有限公司共同编写了《数据基础设施白皮书 2019》，力图从数据应用的现状与问题出发，总结数据基础设施的内涵与技术特征。在研究的过程中我们认识到，目前对数据基础设施的理解还是非常初步的，数据基础设施是涉及经济、技术，乃至社会发展的宏大命题，这本白皮书只是后续研究的一个起点。我们希望未来能够与产业界和各行业专家共同探讨、研究，不断厘清数字经济大背景下数据基础设施的概念与需求，更好的指导技术、产业和应用的发展。

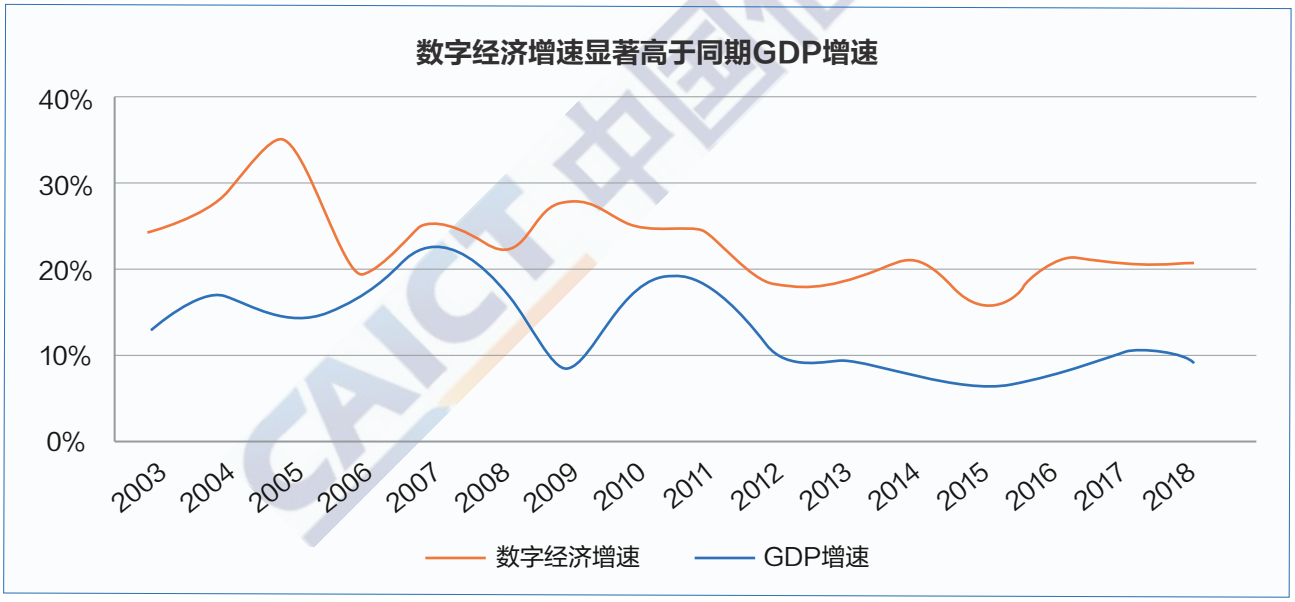
01 数据正在成为数字经济发展关键生产资料

1.1 数字经济蓬勃发展，深刻改变了生产、生活方式

数字经济成为经济发展新动能

受国际经济形势与国内经济结构性改革等因素影响，从2007年起，中国GDP增速从14.2%回落到2015年起的6.9%，经济增速由高速转变为中高速，中国经济进入“新常态”。过去十年，中国数字经济的持续稳定快速发展，成为稳定经济增长的重要途径。2008年，我国数字经济占GDP比重仅为15.22%，2018年我国数字经济规模达31.3万亿元，占GDP比重34.8%，数字经济发展对GDP增长贡献率达到67.9%，超越部分发达国家水平。

2008~2018年，我国数字经济增速显著高于同期GDP增速，并且自2011年以来，数字经济与GDP增速差距有扩大趋势，按照可比口径，2018年我国数字经济名义增长率为20.9%，高于同期GDP名义增速约11.2个百分点。随着数字化加速向传统产业融合渗透，数字经济对经济增长的拉动作用将愈发凸显。



来源：《中国数字经济发展与就业白皮书（2019年）》

数字经济深刻改变了生产、生活方式

移动互联网改变日常生活。中国已经成为全球最大的移动互联网市场，数据显示，截至2018年12月，中国手机网民规模已达8.17亿，用户需求的巨大网络效应带来了一系列广泛的创新，电子商务、网络支付、共享单车、人工智能等新兴领域正迅速重构每一个中国人的生活方式，从而形成日常生活中的数字浪潮。以电子商务为例，十年前中国的零售电商交易额不到全球总额1%，如今占比已超过40%，超过法、德、日、英、美五国的总和。

工业互联网赋能工业企业转型升级。工业互联网通过人、机、物的全面互联，实现全要素、全产业链、全价值链的全面连接，对各类数据进行采集、传输、分析并形成智能反馈，推动形成全新的生产制造和服务体系，优化资源要素配置效率，充分发挥制造装备、工艺和材料的潜能，提高企业生产效率，创造差异化的产品并提供增值服务。以国家电网公司为例，国家电网公司提出“泛在电力物联网”战略，把用户、企业、设备、供应商等人和物全部连起来，实现设备和状态的全面感知，通过把数据汇聚、共享，为用户、电网、供应商等提供数据服务，使数据为社会提供更多价值服务。

以人为本提升社会发展。2017年12月，广东省率先在全国部署“数字政府”改革建设，以数据开放释放“数字红利”，极大提升政府治理能力现代化水平。基于“数字政府”统一基础设施，以数据为核心，盘活政府已有数据中心和社会化数据中心资源，通过数据汇聚、数据治理，建设结构合理、质量可靠的政务“大数据”体系。2018年9月，广东政务服务网正式上线，实现省、市、县、镇、村五级政务服务事项“应上尽上”、“一网通办”，变“群众跑腿”为“数据跑路”。

1.2 数据的有效应用正成为经济社会发展的强劲驱动力

数据是数字经济时代的核心生产要素

社会已经迎来了继农业经济、工业经济之后的数字经济时代，如同农业时代的土地、劳动力，工业时代的技术、资本一样，数据已经成为数字经济时代的生产要素，而且是最核心的生产要素，数据甚至被认为已经超过石油的价值。数据驱动型创新正在向经济社会、科技研发等各个领域扩展，成为国家创新发展的关键形式和重要方向。

数据有效应用推动经济社会发展

各行各业加速数字化进程，对数据的有效应用成为关键。

提高金融风控能力。美国银行2015年的一份调查研究指出，银行每创收100万美元，会平均产生820GB的数据，业务数据量高踞各行业之首，远超紧随其后的电信、保险和能源行业。银行是经营风险的行业，一方面，监管层对银行机构的风控能力提出很高要求，另一方面，风控直接会影响银行机构的利润水平。通过对海量数据的有效利用，能够在用户画像、反欺诈、信用评级等方面大大提高银行机构的效率和风控能力。

提高政府办事效率。以往，群众找政府办事，需要来回跑多次。通过进行数据共享、数据整合，打

破多个部门之间的数据壁垒，来减少人工窗口、缩短审批流程，从而提高办事效率，减少排队等候的情况，更加便民。

扩大企业生产效率。通过数据有效利用能实现企业各业务环节间的信息高度集成和互联，减少不必要的资源浪费。以制造业为例，制造业的研发、采购、物流、生产、库存、销售等环节会产生大量的数据，诸如各工序节拍信息、产品质量信息、发货和收货信息、物料流动信息、客户需求信息、人力资源需求信息等。通过将企业内部和外部各项数据高度集成和互联，能够消除过度生产浪费、等待时间浪费、工序浪费、库存浪费、运输浪费、产品缺陷浪费等，降低生产成本，提高生产效率和产品质量，实现资源优化配置。

提升警务智能化水平。在公安行业建立健全基础数据实时采集、动态更新、高度共享、深度研判的工作机制，汇集来源于公安、政务、社会的数据资源，并面向公安机关及政府部门提供统一的支撑，实现数据资源的交换、集成和服务。通过建立一个以视频图像为主、多种资源关联叠加的视频资源智能化服务体系，打造公安机关视频应用实战的“神兵利器”，全面提升警务智能化水平。

促进经济社会可持续发展。数据的应用有助于推动环保、节能、绿色产业发展，促进环境保护和经济社会可持续发展。例如，利用大数据可以对环境进行立体监测，通过数据模拟技术和排放清单等工具，建立环保大数据系统，提高环境监测数据的可靠性，为经济决策提供科学依据。

1.3 新技术持续推动数据爆发增长

GIV2025报告显示，到2025年，全球将产生180ZB数据。新技术的出现持续推动着数据增长与流动。

4K/8K带来数据存储的需求量激增，以及极致稳定的读写高带宽的需求

当前，信息视频化、视频超高清化已经成为全球信息产业发展的大趋势。从技术演进来看，视频已经从标清、高清进入4K，即将进入8K、AR/VR时代。以广电行业为例，今年3月1日，工业和信息化部、国家广播电视总局、中央广播电视总台联合发布了《超高清视频产业发展行动计划(2019-2022年)》，提出坚持“4K先行、兼顾8K”的总体技术路线，到2022年，中国超高清视频产业总体规模将超过4万亿元。4K超高清的建设和应用，使广电行业IT基础设施在高可靠的基础上，向着高性能、低延迟、集约化的方向转型，尤其对存储平台的能力带来巨大挑战。4K超高清制播业务所产生的数据量比高清多出至少4倍以上，制播的各个环节，如视频剪辑、特效合成、渲染、调色、视频输出等，都需要海量的存储空间以及并发的读写能力。

5G/IoT/车联网带来数据量激增，同时也让数据采集和云边协同能力发生质的变化

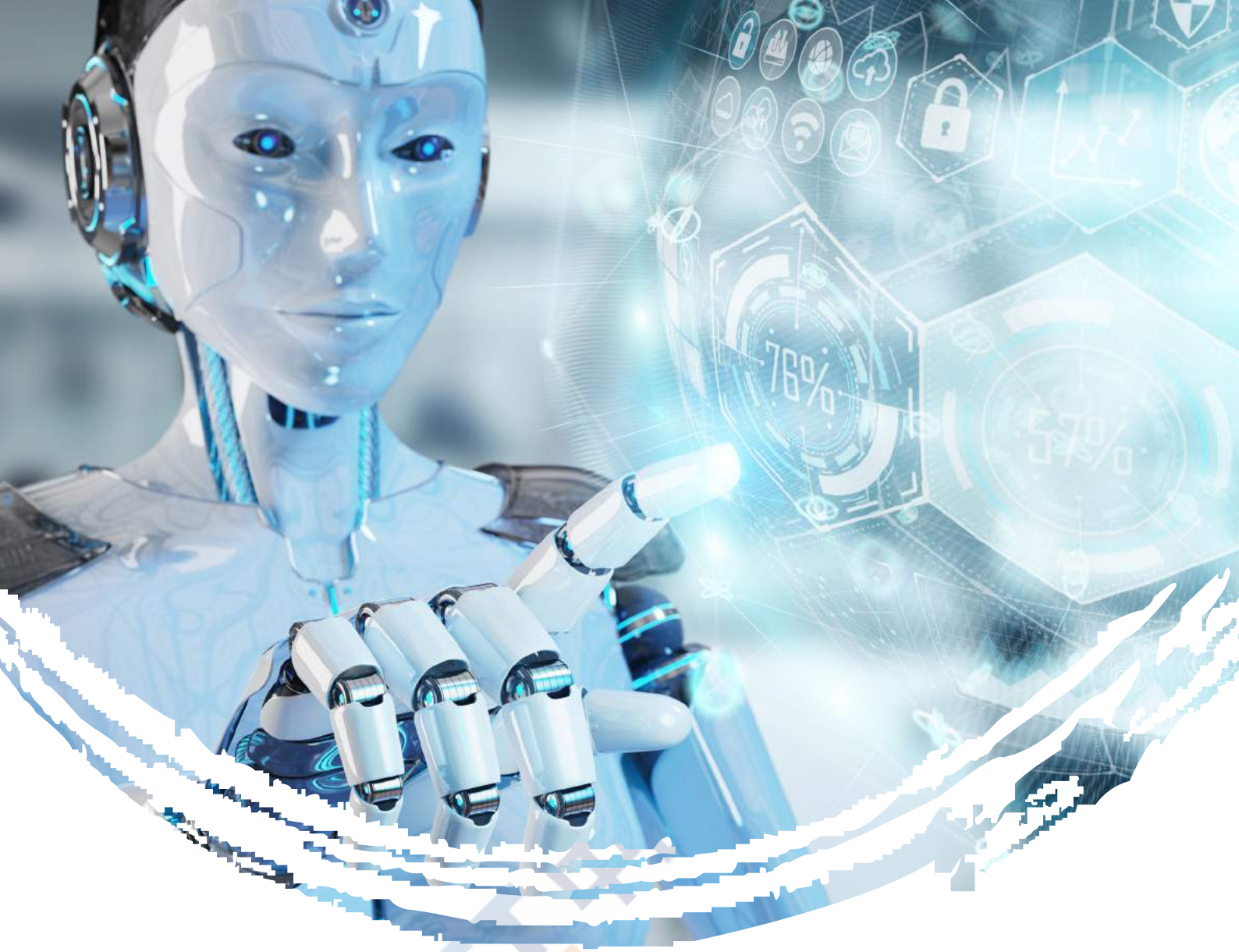
5G通过提升连接速率和降低时延，使得单位时间内产生的数据量急剧增长，单位面积内的联网设备成倍增加，海量原始数据将被收集。4G时代，数据多产生于人与人之间的互联，5G时代，物联网将得到较大程度的发展，人与物、物与物之间的连接将急剧增多，数据采集渠道将更加丰富，如联网汽车、可穿戴设备、机器人等，这也对数据存储与采集技术提出更高要求。同时，5G时代下越来越多的IoT设备将通过边缘计算进行存储、处理和分析，云、边协同能力变得尤为重要。

自动驾驶将产生海量数据，成为新的数据制造机

自动驾驶汽车依赖于安装在车身上的各种传感器传输的大量数据，因此要实现自动驾驶，首先要做好准备迎接海量数据的“洗礼”。在自动驾驶训练时期，以一辆车的信息采集为例，在进行自动驾驶算法路测的过程中，每辆汽车每天将产生60TB的训练数据量，仅仅在2017年，该领域就创造了大约250EB的数据量。2020年前后，自动驾驶汽车将正式上路，每小时将产生4TB的数据，其中包括了关于道路状况、天气、周围物体、交通和街道标志等的实时信息数据，海量数据存储与处理的时代即将到来。

AI/大数据将改变数据的存储周期和形态

首先，AI需要更长的数据存储周期。比如，公安部《公安机关现场执法视音频记录工作规定》明确提出，现场执法视音频资料的保存期限原则上应当不少于六个月，以构筑“更长证据链”。其次，AI需要全数据训练、处理和分析。在数据规模化增长的趋势下，可以按温度来定义不同访问频率的数据：经常被访问的数据称为热数据，而较少被访问的数据称为冷数据，处于中间状态的称为温数据。应用AI之后，需要数据能在冷、温、热之间随时进行切换。



02 数据应用的现状总结

新技术和产业的出现，正在加速企业的数字化转型，大量新的硬件与应用带来数据量快速增长的同时，也让数据类型越来越多样化。生产、采集和保存尽可能多的数据，用于全量分析以洞察先机，成为企业的共识。海量数据蕴含巨大的价值，也给存储系统带来了前所未有的挑战，数据存不下、流不动、用不好成为了各行业数据应用最普遍的难题。

2.1 存不下——数字化浪潮下的海量数据存储挑战

创新业务推动企业的数据量从PB级向EB级迈进，根据《华为全球产业展望GIV》预测，全球新产生的数据量将从2018年的32.5ZB快速增长到2025年的180ZB。由于存储系统仍为传统架构以及成本等原因，当前企业数据仅有不到2%被保存，数据“存不下”的问题日益严重。

- **存储扩展性不足：**传统存储由独立的控制器与硬盘框组成，当容量不足时可增加新的硬盘框进行级联，但由于控制器的处理能力受限，存储的扩展能力非常有限。在政务云建设中，省级平台通常需要规划至PB级的容量，单套存储已经无法满足需求，因此只能部署数十套高端、中端和低端的设备，导致管理的复杂和数据的割裂。

- **存储协议类型单一：**非结构化数据逐步成为企业数据的主体。随着电商、物联网等业务扩张，80%的新增数据由各类音视频、日志等非结构化数据构成。然而传统存储协议类型单一，无法同时满足块、对象、文件、大数据等多样性数据的存取需求，企业不得不为每一种新的数据类型新增一种存储设备，增加了高效利用存储资源的难度。

- **存储成本依然高昂：**越来越多的企业选择将数据长期保存。2017年起，移动运营商因合规性要求，将其设备日志的保存周期从2个月增加至6个月。这意味着其数据存储服务器的设备规模将增加至少2倍。传统的架构中，服务器因存储需求不断扩容，但CPU的使用率却始终处于较低的状态，资源得不到合理利用，无疑会对采购成本和维护成本造成更大的压力。企业不得不因为存储成本而放弃大量宝贵数据。

2.2 流不动——由来已久的数据孤岛难题

孤立的数据价值并不显著，只有当数据像水一样流动起来，才能打破“数据壁垒”，最大化释放其价值。然而当前企业保存下来的数据，由于技术与流动性问题，只有10%的数据能得到分析，数据孤岛、多样性设备、业务迁移成为数据“流不动”的主要瓶颈。

数据的“三类孤岛”

- **应用孤岛：**不同应用产生的数据分别存放在不同的存储系统中，而且这些数据由于各自的特征，彼此之间是无法共享使用的，即形成“应用孤岛”问题；

- **管理孤岛：**为对生产数据加以保护和使用，会将生产数据的一个副本，拷贝到各个系统（如备份、容灾、归档、开发测试和分析系统）中进行管理和使用。即便是同一份数据，为实现不同目的，还需分别存储、管理和使用，即形成“管理孤岛”问题；

- **地理孤岛：**由于企业的更新换代，将存在多套存储设备，比如生产环境、非生产环境、云环境和边缘环境，企业的数据将存放在不同的地方，形成“地理孤岛”问题。

资源的“三堵高墙”

产生上述问题的根本原因：企业在建设数据基础设施时，从满足客户的诉求出发并考虑投资成本问题，会选择不同的计算资源、网络资源和存储资源来分别满足客户的不同诉求。

- **算力墙：** 各个存储系统的CPU能力，仅供本系统使用，无法将算力资源共享使用，形成各存储系统之间的“算力墙”；

- **网络墙：**各个网络都有各自的协议，彼此之间无法互连互通，即各个网络之间形成“网络墙”；

- **介质墙：**存储介质的性能、容量和成本各不相同，客户会选择合适的介质存储数据，这使得数据分别存储在不同系统的不同介质中，而且这些数据很难共享访问，即各个存储介质之间形成“介质墙”。

2.3 用不好——数据供应不足造成应用复杂低效

海量的数据孕育了前所未有的机遇，也带来了巨大的挑战。甚至有人说，从来不缺数据，数据多了反而成为一种负担。也有人说，数据只是资源，而不是资产，很难产生价值。其根本原因是没有用好数据，数据没有释放价值。而影响数据价值释放的主要原因是数据供应不足，无法反馈业务本质，支持业务决策：

大量数据未存储

企业每天会产生大量数据，但传统的数据录入需要预先的人工规划，这导致大量非结构化数据以及一些新型的数据无法进入系统（例如IoT数据、视频数据、图片数据等）。数据的缺失会削弱对业务的感

知，无法真实及时地反映出业务本质。

找不到数据

传统企业通常通过数据表来管理和分析数据，规模较大的公司数据表甚至可以达到数百万张，而且分散在各个业务系统中。如果没有统一数据目录和全局数据视图，要在上百万张报表中找到特定的数据，好比大海捞针，无法应对灵活多变的业务需求。

谁对数据负责

在大数据时代，一个典型的分析业务通常需要跨平台的数据协同。如果已经接入的数据无法满足分析需求，需要从前端多个业务系统获取新的数据，再加上缺乏统一的隐私与安全共享机制，数据就需要经过多部门间协调、拉通、核实才能获得。数据分析的链路冗长，一旦出现问题，就需要“六方会谈”才能定位，无法保证数据供应稳定和高可用，更无法实现高效的数据融合分析。

03 构建数据基础设施迎接变化与挑战

社会数字化、智能化加速发展，海量的数据带来了巨大的挑战，也孕育了前所未有的机遇。各行各业都在加速数字化和智能化进程，越来越多的企业已经意识到，数据基础设施是数字经济成功的关键，而数据“存不下、流不动、用不好”等问题也促成了各行业积极构建新型数据基础设施，加速实现数据价值变现。

3.1 数据基础设施定义

数据基础设施的范围应涵盖接入、存储、计算、管理和数据使能五个领域，通过汇聚各方数据，提供“采-存-算-管-用”全生命周期的支撑能力，构建全方位的数据安全体系，打造开放的数据生态环境，让数据存得了、流得动、用得好，将数据资源转变为数据资产。新的数据基础设施是传统IT基础设施的延伸，以数据为中心，服务于数据，最大化数据价值。

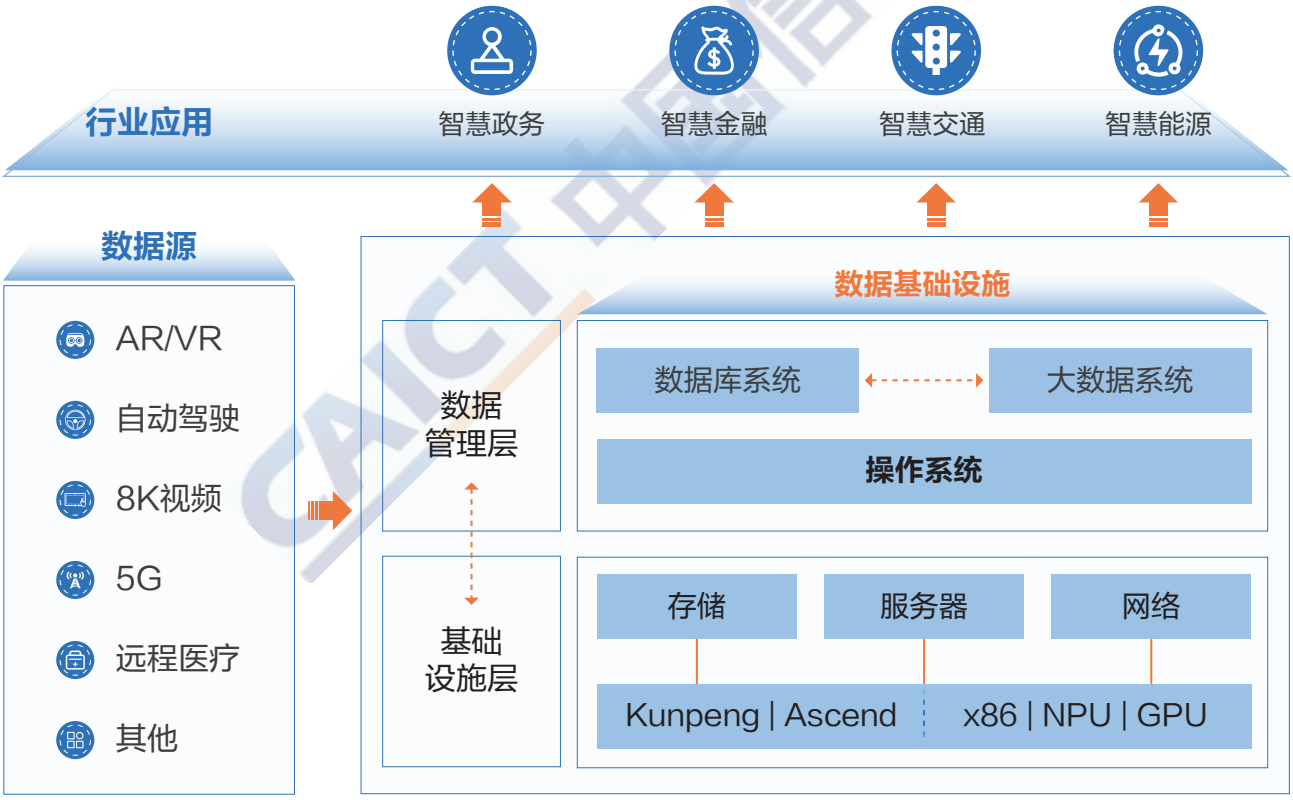


图3-1 数据基础设施

数据基础设施由基础设施层和数据管理层组成，其中基础设施层包括存储、计算、网络等硬件设施，数据管理层由操作系统、数据库系统及大数据系统组成，构成支撑数据存储及数据全生命周期管理的软件设施。

在基础设施层，区别于传统的硬件设施，数据基础设施将引入多样性计算，从单一算力到多样性算力，匹配多样性数据，让计算更高效；存储也会从单一类型存储走向多样性融合存储，构建融合处理基础，应对存储效率低、管理复杂的问题。

在数据管理层，将结合大数据系统和数据库系统提供的“采-存-算-管-用”全流程的软件支撑，从单一处理向多源数据智能协同、融合处理发展，应对更实时和智能的数据应用需求，加速实现数据价值。

数据基础设施需要面向数据构建全方位的安全体系，保障数据端到端的安全和隐私合规，打造开放的数据生态环境，推动全社会数据的共享和开放，创造更大的价值。

3.2 数据基础设施的特征与趋势

数据基础设施应具备以下5个特征：融合、协同、智能、安全、开放，以帮助企业实现存储智能化、管理简单化和数据价值最大化。

3.2.1 融合

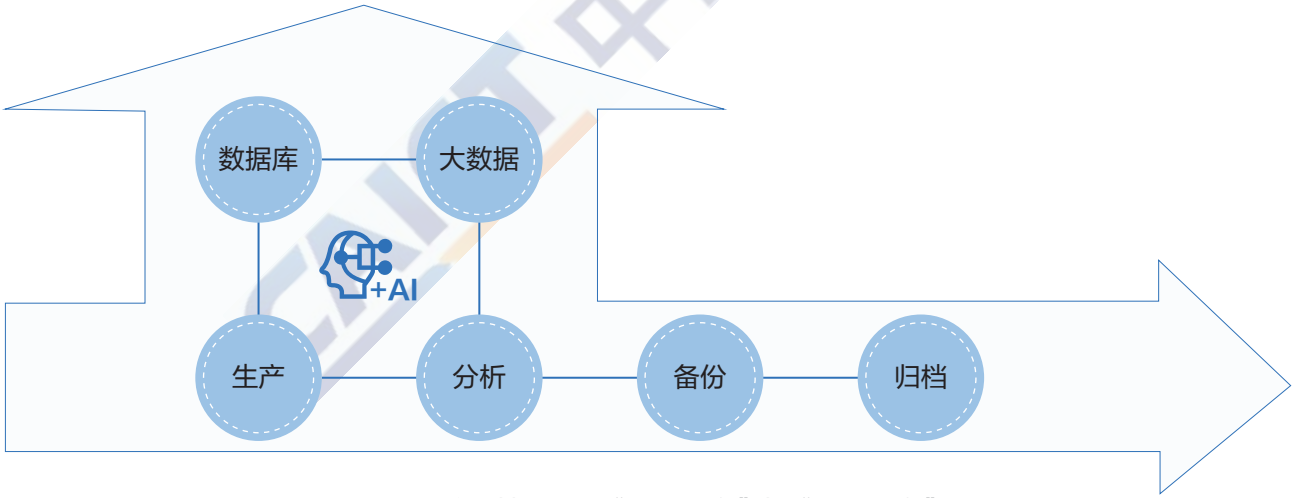


图3-2 数据基础设施“横向融合”与“纵向融合”

数据基础设施正在向“一横一纵”两种融合架构发展。

横向融合是指数据全生命周期存储的融合。数据产生的第一环节是生产存储，以支撑交易型的数据处理；通过扩展至分析型存储来支撑核心的分析业务；备份存储进一步扩展分析场景；主存增加混合云备份、分级等特性，实现冷数据上云。通过对生产存储叠加轻量化备份和管理特性，拓展存储场景，实现从热到冷的数据全生命周期存储整合。

纵向融合是指数据处理与数据存储的垂直优化。包括交易型数据处理与生产存储相融合，提升性能，增强可靠性；数据分析与分析型存储融合，提高分析效率。在存储层，通过重定义存储架构，将块、文件、对象、HDFS等多种存储服务融合，打通数据孤岛，解决多样性数据存储和共享问题；协议方面，通过多协议融合技术，实现一份数据同时支持数据库、大数据、AI等多种业务的分析需求，节省数据无效流动时间，让分析更高效；算力层面，通过将数据库、大数据、AI多引擎融合分析和多样性算力统一调度，降低海量数据处理难度，实现高效分析；管理层面，通过将AI融入存储全生命周期管理，从资源规划、业务发放、系统调优、风险预测、故障定位等方面实现智能运维，从容应对数千节点规模的复杂管理。

数据基础设施五个层面的融合

异构算力融合

随着AI和机器学习的规模使用，数据基础设施必须支持以GPU，FPGA和ARM为代表的异构计算，即从单一算力到多样性算力，匹配多样性数据，为中心、云和边缘提供更高性能的计算资源，使能数据基础设施中的应用更高效运行。

存算融合

存算融合是指将一些数据的处理直接在存储控制器中的盘内进行处理，将计算出来的有效数据返回给计算层，这样可以最大限度地减少在存储层和计算层之间移动的数据量，提升计算效率。

数据库存储融合

数据库存储融合指采用计算-存储分离部署的架构，数据库计算和存储资源可以灵活配置，根据业务需要各自独立进行弹性扩展，使得资源匹配更精准、更合理，大幅提升资源利用率。

协议融合

协议融合是指数据在生命周期中以不同的协议存放在不同的地方，打破这种协议限制，将数据在逻辑上集中，即支持多种应用和数据源的接入，并通过开放式数据接入框架，灵活扩展接入第三方数据源。实现“业务到哪，数据到哪”的全连接，让数据取之有“道”。

格式融合

格式融合是指为上层应用和客户端提供工业界标准接口，能够支持多种服务，如块存储服务、文件存

储服务、对象存储服务和大数据存储服务。消除传统数据基础设施中多类型存储系统烟囱式构建而形成应用孤岛。实现一份数据同时满足数据库、大数据、AI等多种业务的分析需求。

3.2.2 协同

大数据的本质是复杂数据的处理技术，它和成熟的数据库、数据存储技术是相辅相成的关系。因为这种复杂性，在硬件、算力、数据等趋于融合的过程中，多种数据源以及与之相关联的特定的数据处理技术还是长期并存的，需要对异构异地数据源进行协同分析。

数据基础设施的六个协同场景

跨数据源协同分析

实现分散在多个数据源的多张数据表进行交叉分析。如常见的数据源：Hive、SparkSQL、MPPDB、ES、HBase、Oracle、MySQL等。

跨域协同分析

实现分散在异地数据中心的多张数据表进行碰撞分析。可以像访问本地数据表一样访问异地数 据表，并可以将分散在多地的数据表进行碰撞操作。

云边协同分析

目前普遍存在“云—边—端”三级的硬件基础设施。其中，云侧和边侧均有数据持久化存储介质，可用于临时或长期地保存业务数据。实现云边协同分析，即实现云侧和边侧之间特殊的跨源、跨域协同分析。

异地数据即时访问

只要数据进入一个数据中心，通过协同分析联网的其他数据中心就可以马上访问这部分数据，而不需要等待数据复制到本地。

统一访问接口

协同分析对外提供SQL或命令行等统一的查询接口，降低开发人员的学习门槛。开发人员无需关心数据的存储位置，只需要像处理本地数据一样提交处理任务。

跨域计算能力共享

同一个企业组织的多个数据中心分布在不同地域，不同地域之间的计算资源利用往往不平衡，通过

协同分析的任务分发和调度能力，可以实现跨域的计算能力共享，提升整体资源利用率。

协同的关键技术

智能算子下推

当前在跨数据源查询时，无法将算子和计算任务下推至数据源，造成存储节点和计算节点间大量不必要的数据传输，严重影响SQL引擎性能。智能算子下推技术正是为了解决上述的“跨数据源复杂查询传输效率低，耗时长”的业务痛点，提供分布式计算下推能力，将算子和计算任务都下推到数据源，大幅减少从数据源表拉取的结果集，避免不必要的数据传输，提升查询性能。

计算任务下推

在跨数据中心联合查询场景，考虑到数据中心之间的数据安全、网络带宽，及数据中心的集群算力等因素，以数据中心为单位，将查询分解成多个子任务，并下推至对应的数据中心去执行，这样能最大程度降低对网络传输带宽的消耗，提高查询的响应效率。

跨域高速数据传输

跨数据源查询的主要瓶颈在于可用带宽资源不足、查询数据网络传输耗时过长、网络质量不可控导致的查询任务中断、失败。在带宽有限、网络条件差的环境中，实现高速、可靠数据传输和优化，是实现协同分析技术实用化的关键。

3.2.3 智能

数据智能是一个跨学科的研究领域，它结合大规模数据处理、数据挖掘、机器学习、人机交互、可视化等多种技术，从数据中提炼、发掘、获取有揭示性和可操作性的信息，使数据“智能”，为人们在基于数据制定决策或执行任务时提供有效的智能支持。

数据智能的标志是数据驱动决策，让机器具备推理等认知能力，大数据能够指导决策。同时完成业务数据化的进程，开始进入到业务智能化，依靠数据去改变业务。

智能的数据基础设施应该从每个环节都能够提供智能化的能力支撑。

数据基础设施智能化关键环节

智能芯片

按技术架构来看，智能芯片可以分为通用类芯片（CPU、GPU、FPGA）、基于FPGA的半定制化

芯片、全定制化 ASIC 芯片、类脑计算芯片（IBM TrueNorth）等。另外，主要的人工智能处理器还有DPU、BPU、NPU、EPU等适用于不同场景和功能的人工智能芯片。

随着互联网用户量和数据规模的急剧膨胀，人工智能发展对计算性能的要求迫切增长，对CPU计算性能提升的需求超过了摩尔定律的增长速度。同时，受限于技术原因，传统处理器性能也无法按照摩尔定律继续增长，发展下一代智能芯片势在必行。未来的智能芯片主要是在两个方向发展：一是模仿人类大脑结构的芯片，二是量子芯片。智能芯片是人工智能时代的战略制高点，预计到2020年人工智能芯片全球市场规模将突破百亿美元。

智能软件框架

面对海量的数据处理、复杂的知识推理，常规的单机计算模式已经不能支撑。计算模式必须将巨大的计算任务分成小的单机可以承受的计算任务，如云计算、边缘计算、大数据技术提供的基础计算框架。当前人工智能普遍使用通用的开源框架来进行模型的训练，比如：TensorFlow、PyTorch、MxNet、Caffe等。不同的数据使用不同的框架会得到不一样的模型，这些模型最终将用于现实中的推理。

智能数据治理

AI可以解决数据治理的一些痛点：对人工流程的依赖和对专家的依赖。

数据治理需要人工流程保障一系列数据规范、标准的贯彻执行。而智能化的数据治理能够让数据规范和标准的保障自动判断，自动完成，减少对人工审核的依赖；数据治理需要大量数据专家理解数据，理解业务，构建数据安全和数据质量体系，基于AI的数据治理平台通过算法理解数据和业务，对不同的数据自动采取相应的分类安全和质量保障体系，降低对专家的依赖。

数据治理的智能化可以降低客户数据治理方案的总体成本，缩短上线周期，减少对人的依赖。

3.2.4 安全

数据基础设施承载着海量的数据，包括业务的核心数据以及隐私数据。这些数据支撑着企业的所有业务和运营，关系着企业的生命线。需要构建全方位的数据安全体系，帮助企业实现数据在全生命周期过程中的数据不丢失、不泄露、不被篡改、业务永远在线、可追溯和隐私合规。

数据基础设施系统和数据使用方式给安全保护带来了**新的挑战：**

- 海量数据集中后无形中增加了黑客单次攻击获取的收益，降低了攻击成本；

- 分布式计算和存储增加了攻击面和配置管理难度，安全风险更难发现；
- 组件、数据、用户多样，数据误用风险提升；
- 数据流动路径的复杂化导致追踪溯源变得异常困难；
- 数据分析、共享带来新的隐私和合规风险；
- 非结构化数据快速增长，数据全生命周期融合，隐私合规风险激增；

数据基础设施应具备全方位的安全防护体系

数据基础设施应具备平台安全、数据安全、隐私合规三个层面全方位的安全技术体系，打造可信的数据基础设施，帮助企业实现数据在全生命周期过程中的数据永不丢失、不泄露、不被篡改、业务永远在线、可追溯和隐私合规。

平台安全

系统自身的安全和防攻击性是安全防护体系的基石，需要从产品的需求、设计、开发、测试、交付和运维的整个生命周期进行管控，确保系统具备预期交付承诺的安全能力，满足交付质量的要求。基础设施平台安全包括介质、芯片、板卡等硬件设备安全，操作系统、数据库、固件等软件安全，以及网络、协议等安全。

数据安全

是指基础设施为支撑数据存储、传输、处理等全生命周期过程提供的数据安全保护能力，如数据加密、数据隔离、访问控制、完整性校验等。数据融合背景下，由于缺乏有效的安全访问控制，不同网络融合、各种数据汇集，数据泄露及滥用风险成为主要矛盾之一。保障数据的安全，要回答好三个问题：数据在哪里，安不安全；数据去哪里，该不该去；数据谁在用，该不该用。

隐私合规

是指基础设施为保障数据存储、移动、再利用等过程中的合规提供的能力，如数据脱敏、违规分析、密文搜索、同态加密等。欧盟10月4日发布非个人数据移动条例，放宽非个人数据流动限制，以推动欧盟数字经济发展。在该条例下，个人数据的准确识别和数据脱敏将发挥重要的推动作用。二级存储产品将生产业务的备份、复制、归档数据统一存储、统一管理，并及时将副本数据用于开发、测试和数据分析，在这种端到端、多方使用数据的场景下，做好数据的访问控制和脱敏变得尤为重要。

3.2.5 开放

“开放”的数据基础设施需要包容开放的技术和产业生态。

数据产业是一个有众多细分领域、众多参与者的产业，它需要数据、产品和服务间的紧密协同，而数据基础设施作为其中的关键支撑环节，涉及到硬件产业、软件产业，以及各类开源技术、闭源技术等，这就决定了数据基础设施具有生态复杂性，需要很强的生态协同能力，并通过技术和产业的开放性来吸引更多的参与者以保持生态的活力。

构建“平台+生态”的数据基础设施新模式，需要产业各环节的协同操作，包括基础设施和应用服务间的协作、同类型供应商之间的协作、上下游供应商之间的协作、甚至内部产品之间的协作。使产业链上下游实现高效率、低成本的多赢局面。

实现产业生态开放与协同的两个重点

制定公平、透明规则，建立生态信任体系

开放的生态体系中包含了不同的参与者，代表了不同的利益诉求，在缺乏信任的情况下，参与者之间的互动会演化为竞争性的活动。对于开放性的数据和产业生态，建立生态体系内的信任十分重要。开放而有序的生态能为参与者提供发展的“自主权”，并在有需要的时候，提供公平性、透明度的规则来维护参与者的权益。

建立价值分享模式，谋求产业长期发展

开放的平台和生态使产业传统的“分蛋糕”模式逐步转向一起“做大蛋糕”。生态体系中的利益主体通过建立产业链间高效的协同机制，形成良性互动的有机合作关系，以实现产业的持续扩大。未来开放的生态模式将会类似于成熟的软件开源模式，贡献者的名誉将有助于他们在未来的市场化，以获得更多潜在的利益回报。

04 总结与展望

过去十年，中国数字经济蓬勃发展，深刻改变了人们的生产和生活方式。而数据已经成为了数字经济时代的最核心的生产要素。数据驱动型创新正在向经济社会、科技研发等各个领域扩展，成为国家创新发展的关键形式和重要方向。

飞速发展的通信和互联网技术以及随之产生的新型应用需求带来了数据爆发式的增长。海量数据蕴含巨大的价值，在带来更多机遇的同时，也给传统的IT基础设施带来了前所未有的挑战，数据存不下、流不动、用不好成为了各行业数据应用最普遍的难题。培育和建设新的数据基础设施成为了解决这些数据应用问题的关键。

数据基础设施是传统IT基础设施的延伸，以数据为中心，服务于数据，以最大化数据价值。它涵盖数据接入、存储、计算、管理和使能五个领域，提供“采-存-算-管-用”全生命周期的支撑能力。数据基础设施需要具备全方位的数据安全体系，旨在打造开放的数据生态环境，让数据存得了、流得动、用得好，最终将数据资源转变为数据资产。

数据基础设施应具备融合、协同、智能、安全、开放5大特征，以帮助企业实现存储智能化、管理简单化和数据价值最大化。融合指的是“一横一纵”的融合模式，横向融合是数据全生命周期存储的融合，纵向融合是数据处理与数据存储的垂直优化；协同指的是支撑异构异地数据源的协同分析；智能指的是贯穿数据基础设施每个环节的智能化的能力支撑；安全指的是提供平台安全、数据安全、隐私合规全方位的安全防护体系；开放指的是数据基础设施的发展需要包容开放的技术和产业生态。

企业向数据驱动型企业转型的过程不是一蹴而就的。随着企业在每个阶段对自身数据认知的不断提升，其对基础设施（包括数据基础设施）的要求也会逐步提升。成长中的数据基础设施，其稳定性和先进性会深刻影响到企业数字化转型的效果和进程。未来，打造开放的产业生态也是数据基础设施发展的关键要素。

中国信息通信研究院

地址：北京市海淀区花园北路52号

邮政编码：100191

联系电话：010-62304839

传真：010-62304980

网址：www.caict.ac.cn

华为技术有限公司

地址：深圳龙岗区坂田华为基地

邮政编码：518129

联系电话：+86 755 28780808

网址：www.huawei.com